*Published online 16 November 2011* *Nucleic Acids Research, 2012, Vol. 40, Database issue* **D627–D631**
*doi:10.1093/nar/gkr1020*

# *ProOpDB*: *Pro*karyotic *Op*eron *D*ata*B*ase

**Blanca Taboada[1], Ricardo Ciria[2], Cristian E. Martinez-Guerrero[2] and Enrique Merino[2,*]**

[1]Centro de Ciencias Aplicadas y Desarrollo Tecnológico, Universidad Nacional Autónoma de México, México, D.F. and [2]Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

Received August 13, 2011; Revised October 19, 2011; Accepted October 21, 2011

## ABSTRACT

**The *Prokaryotic Operon DataBase* (*ProOpDB,* http://operons.ibt.unam.mx/OperonPredictor) constitutes one of the most precise and complete repositories of operon predictions now available. Using our novel and highly accurate operon identification algorithm, we have predicted the operon structures of more than 1200 prokaryotic genomes. *ProOpDB* offers diverse alternatives by which a set of operon predictions can be retrieved including: (i) organism name, (ii) metabolic pathways, as defined by the *KEGG* database, (iii) gene orthology, as defined by the *COG* database, (iv) conserved protein domains, as defined by the *Pfam* database, (v) reference gene and (vi) reference operon, among others. In order to limit the operon output to non-redundant organisms, *ProOpDB* offers an efficient method to select the most representative organisms based on a precompiled phylogenetic distances matrix. In addition, the *ProOpDB* operon predictions are used directly as the input data of our Gene Context Tool to visualize their genomic context and retrieve the sequence of their corresponding 5′ regulatory regions, as well as the nucleotide or amino acid sequences of their genes.**

## INTRODUCTION

Recent developments in sequencing methodologies have tremendously increased the repertory and size of genome sequence databases. More than 1200 prokaryotic and eukaryotic genomes have been completely sequenced now, and the sequences of many others are close to being finished (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html). Moreover, this trend is expected to continue as new and more efficient sequencing techniques are developed. In this scenario, it becomes essential to develop new and better predictive tools for characterizing the properties of sequenced genomes. One of these properties that have been subject to bioinformatics studies is the tendency for coordinating the expression of metabolically or functionally related genes. In prokaryotic genomes, these genes are commonly found contiguously arranged on the same transcriptional strand and are co-transcribed in the same transcription units, called operons. Operons are the basis to determine structures of a higher level of genomic organization as well as different cellular functions, providing important insights for experimental designs. Consequently, diverse computer methods for the identification of operons have been developed and used to predict operons in model organisms, such as *Escherichia coli* (1) or *Bacillus subtilis* (2) or in the fast growing set of fully sequenced genomes. As a result of this work, important databases with operon predictions in prokaryotic genomes have been developed and are publicly available. The strengths and characteristics of each of these different databases vary from one to another. For example, *DOOR* (Database of prOkaryotic OpeRons) (3) offers diverse querying methods to find particular operons, including those with RNA genes. In addition, this database provides similarity scores between operons by which related operons in different organisms can be retrieved. *DOOR* can also identify over-represented sequence motifs in regulatory regions of the selected operons using MEME (4) or CUBIC, a motif identification program developed by the authors. A second database is *MicrobesOnline* (5) which is one of the most complete databases designed to integrate functional genomic data with comparative genome analyses. In order to accomplish this goal, *MicrobesOnline* has two main approaches, the phylogenetic approach, including a tree-based browser and tools for users to build their own trees, and the functional approach, including a wide set of tools to analyze microarray gene expression data and find genes that are co-expressed or have a particular expression profile. *MicrobesOnline* also provides tools to identify conserved regulatory motifs. Another database public available is *OperonDB* (6) which has the most updated list of operons predictions including 1059 bacterial genomes. Finally, *ODB* (Operon DataBase) (7) aims to collect operons that have been experimentally determined or are conserved in different organisms to define a set of reference operons. In this database, operon predictions in

*To whom correspondence should be addressed. Tel: +52 777 3291634; Fax: +52 777 3172388; Email: merino@ibt.unam.mx

a genome are accomplished by mapping orthologs genes on the set of reference operons. This database provides graphical capabilities to inspect the gene context of the selected operons.

The operon accuracies of the computer algorithms used in the mentioned databases, vary from ~80% [in the cases of *OperonDB* (6) and *MicrobesOnline* (5,8)], to ~90% [in the case of the *DOOR* database (3,9)], when the predictions are made on model organisms, such as *E. coli* or *B. subtilis,* and the training and testing datasets are from the same organisms. Nevertheless, it is common to observe an important accuracy decrement when datasets belong to different organisms. In this sense, our *Pro*karyotic *Op*eron *D*ata*B*ase (*ProOpDB*) uses a novel operon prediction algorithm with one of the highest accuracy levels ever reported (10) regardless the source of training and testing datasets. In addition, *ProOpDB* offers diverse alternatives to retrieve a specific set of operons, making it unique in its kind. *ProOpDB* is one of the most precise and complete repositories of operon predictions now available. It includes >1200 prokaryotic genomes and a total of 2 549 412 predicted operons, including RNA genes.

## MAIN ATTRIBUTES OF *ProOpDB*

As previously mentioned, there are several kinds of operon databases that offer different advantages to the users. The main attributes of *ProOpDB* are as follows:

### *ProOpDB* contains the most accurate bacterial operon predictions

The operon predictions accuracy of *ProOpDB* is one of the most important characteristics of our database. These predictions were generated by our recently published operon identification neural network method, which was successfully tested on the set of experimentally defined operons of *E. coli* and *B. subtilis*, with accuracies of 94.6% and 93.3%, respectively (10). As far as we know, these are the highest accuracies ever obtained when predicting bacterial operons. Furthermore, one fundamental advantage of *ProOpDB* over other operon databases is that the performance of the algorithm used to predict operons remains outstandingly high even when the training and testing organisms are not the same. For example, when our algorithm was trained with *B. subtilis* data to predict *E. coli* operons the accuracy obtained was 91.5%, and when the training procedure was done with *E. coli* data to predict *B. subtilis* operons, the accuracy was 93% (10). We consider that these accuracies are significantly high, especially taking into account that the highest accuracy previously reported in a similar analysis was only of 83% (9). Furthermore, to evaluate the performance of our operon predictive method in organisms different than *E. coli* and *B. subtilis*, we tested it on a set of 202 experimentally determined operons compiled in the ODB database (7) that includes 433 operonic-gene pairs in 50 partially sequenced genomes. The accuracy reached by our method in this data set was 92.4%. In addition, we also tested our method in a set of 1145 operonic gene-pairs of 522 predicted operons from a genome-wide transcriptional study (11). In this case, the accuracy was also very high 91.3%. These results show the potential of our method to accurately predict the operons of any other newly sequenced organism. This generalization capability is archived since our operon predictions are based on the functional relationships of contiguous genes defined by the *STRING* database (12) that integrates the information from distinct kinds of sources of different organisms, such as gene neighborhood, gene fusion, gene co-occurrence, gene co-expression and protein–protein interactions.

### Retrieval of operons in *ProOpDB* can be based on metabolic pathways

In addition to the commonly used operon criteria, such as gene name or gene ID in specific genomes, *ProOpDB* allows operon retrieval and visualization by specific metabolic pathways of cellular processes as defined by the well-known *KEGG* database (Kyoto Encyclopedia of Genes and Genomes) (13). This retrieval mode based on metabolic pathways allows any user to perform the operon analysis of an organism, or group of organisms, from a more integrated point of view in accordance with their cellular and biochemical knowledge. For instance, if a user is interested in knowing the operons of *E. coli* K-12 MG1655 involved in chemotaxis, he/she can simply specify the *KEGG* pathway 02030 to find out that, in this organism, there are 20 genes clustered into 9 operons that are involved in this cellular process. Furthermore, with *ProOpDB* any user can easily identify regulatory motifs, not only related to a specific gene family, but also to a particular metabolic pathway. For example, if a user is interested in identifying regulatory elements of genes related to the thiamine metabolism, it would be easy to retrieve the intergenic sequences of operons encoding genes belonging to this pathway, which corresponds to the *KEGG* pathway 00730. Once that the sequences have been retrieved, they can be used directly as the input data to the motifs identification program *MEME* (4), which has been locally installed in our web server, or as input data to any other motifs identification web servers. After this analysis, the user would be able to identify the conserved sequence motif of the Thi-box riboswitch (14) (Figure 1).

### *ProOpDB* allows the operon retrieval and visualization based on COGs groups

A frequent approach when looking for related operons in databases, such as *DOOR* (3), is to consider the orthology relationships of genes based on the *Cluster of Orthologous Groups* of the *COG* database (16). However, since the set of organisms that has been annotated in the *COG* database is limited to only some of them, the recovery of genes by *COGs* can be highly restricted to certain genomes. To overcome this problem, we have assigned to each gene a corresponding *COG* group using Hidden Markov Models and the *HMMER* program (17), so that, operons in *ProOpDB* can be efficiently retrieved by this criteria.

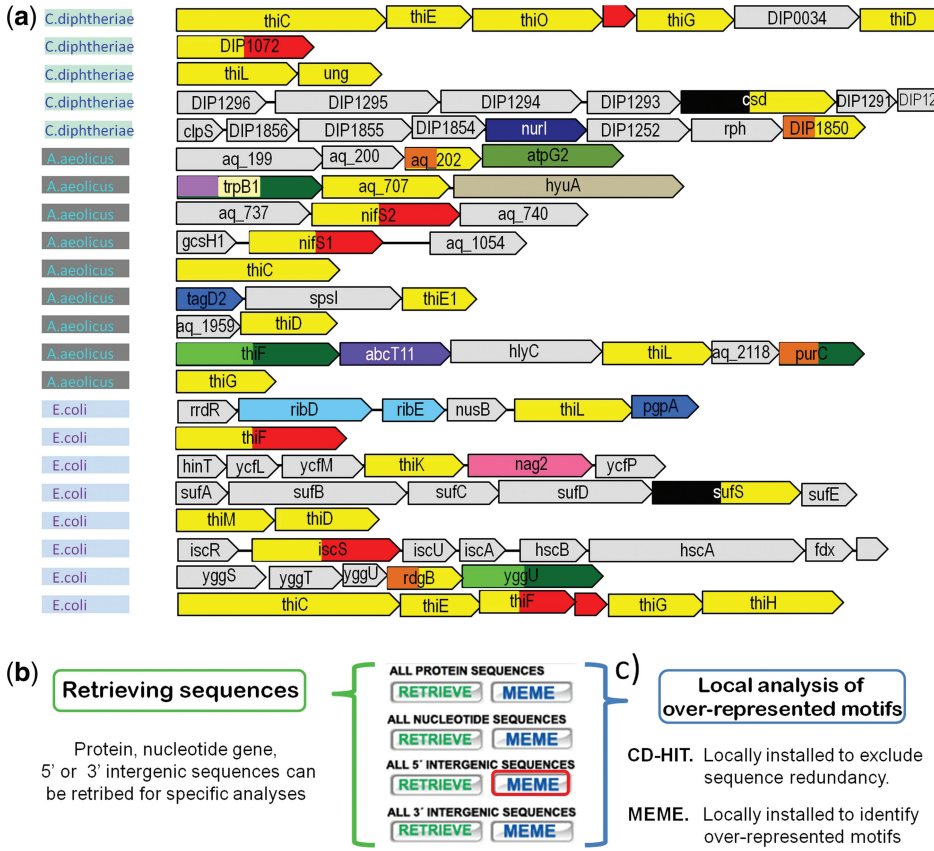


**Figure 1.** Operon structures of genes participating in the thiamine metabolism pathway, KEGG 00730 in different organisms. Among the diverse alternatives offered by *ProOpDB*, the selection based on KEGG pathways allows the comparison of the different transcription units that belongs to a specific metabolic process in different organisms. (**a**) The great diversity of operon organization that is involved in the thiamine metabolism can be observed. It is important to note that genes, in the *ProOpdB* output, are colored in accordance to the feature (phylogenetic—COG, metabolic—KEGG or conserved protein domains—Pfam), that was used to in the operon retrieval process. In our example, the genes are colored based on the KEGG pathway annotations, thus the potential relationships between metabolic pathways can be inferred. For example, genes that belong to the thiamine metabolism (KEGG 00730, yellow color) are part of operons co-transcribing genes of the sulfur relay system (KEGG 04122, red color) and with genes of the purine metabolism (KEGG 00230, orange color) in *Aquifex aeolicus* (Aquificae), *Corynebacterium diphtheriae gravis* (Actinobacteria) and *E. coli* K-12 MG1655 (Proteobacteria). (**b**) The 5′ and 3′ regulatory sequences of the operon as well as the protein and nucleotide sequence of the genes can be retrieved for specific analyses by particular user programs. (**c**) Finger-print analyses can be performed using the locally installed programs in the *ProOpDB* web server. The redundant sequences are eliminated using the CD-HIT program (15) prior the analysis of over-represented motifs using the MEME program (4).

### Operons in *ProOpDB* can be selected based on the conserved domains of their proteins

The conserved domains defined in *Pfam-A* (18) of each gene in *ProOpDB* have been annotated using the *hmmpfam* program of the *HMMER* package (17) so that operons may also be retrieved by a given conserved domain of their corresponding genes. For example, if a user is interested in finding the structure of the operons encoding regulatory proteins with the helix–turn–helix domain of the LysR family (*Pfam* HTH_1), he/she will easily find that almost all of them are monocistronic. In another instance, if the user needs to know the structure of operons carrying the RelB toxin–antitoxin system, using as an input the name of the *Pfam* family *relB*, he/she will discovers that it corresponds to bi-cistronic operons encoding proteins with the conserved domains of RelB and the plasmid stabilization system protein.

### Selection of operons in *ProOpDB* can be made on the basis of a reference gene or a reference operon

An important feature of *ProOpDB* is its facility to show the structures of the operons that contain a gene, or a family of genes, of a particular interest. To this end, we have determined the orthology relationship of genes by the Bi-Directional-Best-Hit criteria using Blast searches, thus all the operons, in a selected set of organisms, containing the orthologs of a reference gene will be displayed. *ProOpDB* can also perform a similar task using a reference operon as input. In this case, all the orthologs of any of the genes of the reference operon will be considered.

### In *ProOpDB* the nucleotide or protein sequences associated to selected operons can be easily retrieved

*ProOpDB* also offers the possibility to download the set of 5′ intergenic sequences of the selected operons. These sequences can be further used to perform fingerprint

analysis to identify regulatory motifs or any other kind of analysis. In addition, it is also possible to retrieve the amino acid sequences of the operons encoded proteins. These sequences are exported as flat files in *Fasta* format.

### The selection of organisms in *ProOpDB* can be done based on their taxonomic order or by their phylogenetic distances

As the number of available sequenced genomes increases, the need of an efficient protocol to select non-redundant organisms from a specific *taxon* or different *taxa* becomes an important issue. For example, in the set of fully sequenced genomes, there are more than 30 *E. coli* strains. In this regard, we have evaluated the phylogenetic distance between every pair of organisms in *ProOpDB,* based on the sequence alignment of the gene concatenation of 31 orthologs present in 191 sequenced genomes that have been selected according to Ciccarelli *et al.* (19). This phylogenetic distance information is used by *PropODB* to select the set of less redundant organisms from a list of all possible organisms in a given *taxon* or *taxa* chosen by the author. This property of *ProOpDB* is particularly useful for identifying regulatory sites in a finger-print analysis. For example, if the user is interested in identifying, in the Gama-proteobacteria, the tryptophan repressor binding site that is used for its autoregulation, he/she could restrict the analysis to this particular phylogenetic group. A search in *ProOpDB* using '*trpR'* as keyword will result in 95 *trpR* genes with this name. To avoid the inclusion of many *E. coli* closely related strains (e.g. *E. coli* 0127, *E. coli* 55989, *E. coli* APEC, *E. coli* BW2952, etc.), the user could ask for a smaller number of organisms, for example 30. Using a pre-compiled phylogenetic distances matrix between organisms, *ProOpDB* will select the less redundant set of these 30 Gama-proteobacteria organisms from where the *trp* regulatory regions can be obtained to be used in the regulatory finger-print analysis.

## IMPLEMENTATION AND WEB INTERFACE

*ProOpDB* is organized in five modules. The first module, *data acquisition*, is dedicated to retrieving primary information from the *KEGG* flat files database (13), including the genomic sequence of organisms and information of their corresponding genes, such as their names, functions and their corresponding metabolic pathways. In the second module, *data analysis,* every gene in *ProOpDB* is annotated with its corresponding COG (16) or ROC (10) groups using our Hidden Markov Models and the *hmmsearch* program of the *HMMER* package (17). In addition, the exact position of conserved protein domains using the *Pfam-A* models (18) was determined using the *hmmpfam* program of the *HMMER* package (17). In this module, BLAST comparisons (20) are performed to identify bi-directional best hits (BDBH) among proteins of the same COG to establish likely orthology relationships. In the third module, *operon predictions*, the operonic or non-operonic nature of every pair of genes is predicted using our recently developed program (10) based on an artificial neural network. The input variables of this neural network are intergenic distances and functional relationships between the protein products of contiguous genes, as defined by STRING database (12), afterward, the structure of all the operons are determined. In the fourth module, *database management*, all the acquired or generated information is stored using a relational database management system (MySQL, http://www.mysql.com/). The fifth module, *web server service*, is related to subroutines and modules required for the accessibility and display capabilities of our web interface. In this module, the names of the genes that were obtained as result of the operon predictions analysis are sent to our web server Gene Context Tool (*GeConT*) for their graphic representation. *GeConT* uses a collection of Perl-CGI programs using the open source code GD graphics library and JavaScript codes to create HTML files.

## FUTURE PLANS

The next update of *ProOpDB* will have a fully automatic update module, so that all of the above-mentioned modules will run automatically. This enhancement will ensure that *ProOpDB* will always be updated.

## ACKNOWLEDGEMENTS

We thank S. Ainsworth for bibliographical assistance, and A. Ocadiz, J.M. Hurtado and A. Martinez for computer support.

## FUNDING

This work was supported by Consejo Nacional de Ciencia y Tecnología CONACyT (grant number 154817) and Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica PAPIIT (grant number 203211). Funding for open access charge: PAPIIT (grant number 203211).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
2. De Hoon,M.J., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2004) Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pac. Symp. Biocomput.*, **9**, 276–287.
3. Mao,F., Dam,P., Chou,J., Olman,V. and Xu,Y. (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**, D459–D463.
4. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
5. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.

6. Pertea,M., Ayanbule,K., Smedinghoff,M. and Salzberg,S.L. (2009) OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res.*, **37**, D479–D482.
7. Okuda,S. and Yoshizawa,A.C. (2011) ODB: a database for operon organizations, 2011 update. *Nucleic Acids Res.*, **39**, D552–D555.
8. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
9. Dam,P., Olman,V., Harris,K., Su,Z. and Xu,Y. (2007) Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res.*, **35**, 288–298.
10. Taboada,B., Verde,C. and Merino,E. (2010) High accuracy operon prediction method based on STRING database scores. *Nucleic Acids Res.*, **38**, e130.
11. Toledo-Arana,A., Dussurget,O., Nikitas,G., Sesto,N., Guet-Revillet,H., Balestrino,D., Loh,E., Gripenland,J., Tiensuu,T., Vaitkevicius,K. *et al.* (2009) The Listeria transcriptional landscape from saprophytism to virulence. *Nature*, **459**, 950–956.
12. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
13. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
14. Serganov,A., Polonskaia,A., Phan,A.T., Breaker,R.R. and Patel,D.J. (2006) Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, **441**, 1167–1171.
15. Huang,Y., Niu,B.F., Gao,Y., Fu,L.M. and Li,W.Z. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
16. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
17. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
18. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
19. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
20. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.